Middle East Journal of Applied Science & Technology (MEJAST)
Volume 8, Issue 1, Pages 01-20, January-March 2025

Performance Evaluation of Hepatitis Disease Prediction in Early Stage Using Machine Learning Techniques

Mst. Sumaiya Akter Mim[1], Md. Julker Nayeem[2], Sohel Rana[3] & Md. Rabiul Islam[4*]

[1,4]*Department of Computer Science & Engineering, Pundra University of Science & Technology, Bogura-5800, Bangladesh.* [2]*Department of Computer Science & Engineering, International Islami University of Science and Technology Bangladesh, Dhaka-1349, Bangladesh.* [3]*Department of Information & Communication Technology, Chandpur Science and Technology University, Chandpur-3600, Bangladesh.*
Corresponding Author (Md. Rabiul Islam) Email: mdrabiulislam521@gmail.com[*]

DOI: https://doi.org/10.46431/MEJAST.2025.8101

Article Received: 07 November 2024 *Article Accepted: 15 January 2025* *Article Published: 23 January 2025*

ABSTRACT

The application of classification approaches utilizing multi-variable with machine learning methods holds immense implications, particularly in the realm of healthcare and disease prediction. Accurate classification of medical conditions, such as hepatitis, is critical for early diagnosis and timely intervention. In order to identify people based on important hepatitis-related characteristics, this study applies advanced machine learning with statistical techniques. It also examines a real dataset in order to create a reliable early detection predictive model. Through this model, we aspire to raise awareness and guide affected individuals toward timely treatment. The paper focuses on comprehensive data preprocessing, including outlier removal, handling class imbalance problem, missing values and extract highly correlated features in order to improve model performance. In our research paper, we have applied mean/mode imputation technique to deal with missing values. Furthermore, we have used z score approach to detect and remove outliers from out dataset and handle class imbalance problem by using oversampling technique. To identify features that are highly correlated, we have used the embedded feature selection approach in our paper. Classic machine learning algorithms, notably K-Nearest Neighbors (KNN), Naive Bayes (NB) and Random Forest (RF) have employed to predict either a person is affected by hepatitis disease or not. To assess the efficacy of our model, we have utilized the 10-fold cross validation procedure. At 97.44%, we have the highest classification accuracy from RF, with Precession, Recall, F1 score and ROC values of, respectively, 0.99, 0.96, 0.97 and 1.00.

Keywords: Hepatitis; Missing values; Class imbalance problem; Early-stage prediction; Machine learning; KNN; NB; RF; Classification.

1. Introduction

Hepatitis is an inflammation of the liver and represents a significant global health challenge affecting millions of individuals [1]. The disease progresses through two distinct phases: acute hepatitis, which is characterized by a short-term infection usually within the first six months, and chronic hepatitis, a long-term ailment that manifests after this initial period. Chronic hepatitis can lead to severe liver damage, altering liver function, potentially resulting in life-threatening conditions. During a hepatitis infection, the immune system responds by releasing inflammatory substances, prompting the liver to generate fibrous proteins like collagen to repair damaged tissues. However, excessive fibrous tissue buildup can hinder blood flow through the liver, ultimately impairing its functionality. Over time, this damage can lead to the death of liver cells, disrupting normal liver operation. Global interest in artificial intelligence (AI) and machine learning (ML) has increased recently, with medical applications leading the way. Automated diagnostic processes and individual health monitoring are just a few examples of how AI is transforming healthcare. Particularly, AI's potential to automate aspects of medication prescription offers significant time savings for medical professionals, enabling them to focus on non-automatable tasks. A key component of machine learning is classification, which is essential for predicting possible classes of data objects. Machine learning algorithms assimilate data to construct models, allowing for intelligent decision-making. Researchers have extensively explored machine learning algorithms to create prediction models based on clinical records, notably in healthcare contexts like diagnosing the presence of hepatitis based on clinical and biochemical data. The liver is vital, and hepatitis threatens it. Early detection is crucial. Data mining in healthcare is key, handling vast data for informed decisions. One of the most difficult tasks facing medical research today is

identifying hepatitis early in a patient's body. As the medical industry witnesses an exponential growth in health-related data, efficient management and analysis of this substantial data have become imperative. Data mining, a subset of machine learning, emerges as a potent tool capable of handling vast datasets and efficiently solving complex problems. This field empowers researchers to extract meaningful insights and make informed decisions from extensive databases. Due to its efficiency and effectiveness, data mining is the preferred choice of researchers for tackling real-world problems, especially in the domain of healthcare and disease detection. The incorporation of ML approaches further enhances the accuracy and efficacy of the analysis, allowing for more precise detection and proactive management of hepatitis, a critical concern in public health. This research attempts to determine which classifier is most effective in predicting the existence of hepatitis in the human body by utilizing different classification algorithms. To deal with the dataset's missing values, the mean and mode imputation approach is applied. Additionally, the boxplot and interquartile range (IQR) have been used to eliminate outliers and the oversampling approach used to distribute the classes equally. Next, we applied the embedded feature selection strategy to find highly correlated features that helped to improve classification performance. Moreover, our research indicates that the RF classifier may be utilized instead of the commonly used NB and KNN.

1.1. Study Objectives

The following are the objectives that this study makes:

(i) Data preprocessing: This work is primarily noteworthy for its steps in data preparation and classification, wherein it applies the mean and mode imputation technique to deal with missing values, the z score technique to remove outliers and the oversampling approach to address issues arising from the unequal class distribution of the dataset.

(ii) Influence of highly correlated features in classification: To get characteristics that are highly related, we have used the embedded feature selection approach. Features with high correlations aid in improving the performance of classification models. In our experimental evaluation, at first, we have calculated classification accuracy of our classifiers by using the dataset before applying our mentioned feature selection approach. Secondly, we have also calculated classification accuracy by using the dataset where exists only highly correlated features. Finally, we have compared these two steps performances. Furthermore, we have also compared the performance of our best classifier with existing research.

(iii) Performance evaluations: Accuracy, Precision, Recall, F1-score, ROC (Receiver Operating Characteristic) and Area Under the Curve (AUC) are the metrics used in the experimental assessment to compare and assess the performances of our mentioned classifiers. These metrics provided precise insights into the efficacy and dependability of our predictive models for accurately detecting hepatitis.

(iv) Comparison of classifiers: The comparison of classifiers involved a meticulous assessment of several ML algorithms, including KNN, NB and RF. Each algorithm has evaluated based on crucial factors such as accuracy, precision, recall and F1-score and roc to gauge their performance in classifying hepatitis. When combined with z-score, oversampling, mean mode imputation, and embedded feature selection algorithms, the Random Forest classification approach outperforms the other chosen classifiers.

(v) Statistical evaluation: ROC curves are created by plotting the True Positive Rate (TPR) vs False Positive Rate (FPR) at different threshold values in order to assess the performance of the classifiers and evaluate the proposed machine learning based system.

The remaining part of our research is outlined as follows. The use of ML algorithms in the prediction of hepatitis disease is reviewed in Section 2. The materials and study framework, including data sources, variable descriptive statistics, data preparation steps, and several classification strategies, are covered in Section 3. This section also mentions the performance measures. In addition to presenting the experimental data, Section 4 illustrates statistical and graphical performances. Section 5 offers a succinct explanation, while Section 6 presents the work's conclusion.

2. Related Works and Motivations

In recent research focused on hepatitis classification, a multitude of machine learning (ML) algorithms have been applied to various datasets, leveraging clinical, biochemical, and demographic features. These algorithms have played a crucial role in predicting and diagnosing hepatitis accurately. H.M. Farghaly et al. [2] developed a comprehensive machine learning (ML) framework focused on diagnosing Hepatitis C Virus (HCV) disease, particularly among Healthcare Workers (HCWs). The study employed several prominent classifiers, including Naive Bayes (NB), Random Forest (RF), Logistic Regression (LR) and K-Nearest Neighbors (KNN). The authors succeeded in early diagnosis of Hepatitis C among healthcare workers using machine learning, addressing occupational risk. However, a limitation of this work is its exclusive focus on predicting Hepatitis C and its application only within the HCW demographic. S. Gundogdu [3] made a significant contribution by developing a robust Support Vector Machine (SVM) model utilizing Principal Components Analysis (PCA) for the detection of Hepatitis C Virus (HCV) disease. The SVM model was designed to be highly effective. However, a drawback of the study was the relatively small sample size utilized for the analysis. Furthermore, the model was primarily focused on predicting HCV, and it was noted that the new components generated by PCA were challenging to interpret, adding a layer of complexity to the analysis. M.M. Majzoobi et al. [4] performed a comprehensive comparison of traditional and ensemble learning methods was undertaken to predict Hepatitis B Virus (HBV) and Hepatitis C Virus (HCV). The study made a valuable contribution by evaluating various methodologies, including Bagging, AdaBoost, Random Forest (RF) and Logistic Regression (LR). However, a limitation of the study was the utilization of a relatively small sample size. Additionally, the study overlooked considering an important aspect-the risk factors associated with hepatitis. M.J. Nayeem et al. [5] performed an insightful comparison of five diverse machine learning techniques for hepatitis prediction, namely KNN, NB, SVM, MLP and RF. An important aspect of the study was the evaluation of prediction outcomes based on different risk factors associated with the dataset. However, a limitation of the study was the presence of an imbalanced dataset, which could impact the accuracy and reliability of the predictions. M.B. Butt et al. [6] focused on utilizing Neural Networks (NN) and specifically the Back-propagation algorithm to predict the stage of hepatitis C. The study made a significant contribution by successfully employing NN in this context. However, a drawback was observed during validation where the precision of the predictions decreased, suggesting a need for further refinement and improvement in the model to enhance its accuracy and reliability. Six machine learning approaches, including LR, DT, RF, KNN and SVM,

were used by M.A. Hafeez et al. [7]. The primary contribution of this research was the identification of the most effective model for predicting HCV disease based on selected features. However, a notable drawback was observed in terms of the implementation being confined to the same package, implying a need for diversifying the implementation for broader applicability and robustness of the model. T.I. Trishna et al. [8], utilized various classification techniques such as Naive Bayes (NB), K-Nearest Neighbors (KNN) and Random Forest (RF) for an in-depth analysis of a hepatitis dataset. The primary objective was to enhance the accuracy of result prediction for each case of data. However, a challenge encountered was the imbalance in the dataset, which can affect the accuracy and reliability of the predictions. Addressing this issue of imbalanced data is crucial for achieving more accurate and meaningful results. S. Hashim et al. [9] used Decision Tree learning Algorithm in order to predict advanced liver fibrosis. R.Y. Krishnabayu et al. [10] used Random Forest technique to Hepatitis disease. The major objectives of this research are to solve imbalance problem and find out highly correlated features. The authors used RFE (Recursive Feature Elimination) and SMOTE technique in their research paper in order to get highly correlated features and solve class imbalance problem respectively. P. Dutta et al. [11] used HGARF technique to improve missing value imputation. R.K. Sachdeva et al. [12] incorporated a systemic method in order to predict hepatitis using machine learning. M.D. Genemo [13] proposed a deep learning-based technique to identify hepatitis and achieved 93.4% accuracy. I.I. Ahmed et al. [14] predicted hepatitis disease by using different machine learning techniques and achieved highest accuracy from DT and RF. The best accuracy for hepatitis C prediction was obtained by SVM and XGBoost, according to a comparative study of many ML algorithms conducted by A. Alizargar et al. [15]. A.Q. Md et al. [16] used ensemble techniques in order to predict liver disease. A. Alotaibi et al. [17] applied ensemble-based machine learning models to predict cirrhosis in hepatitis patients. E. Dritsas & M. Trigka [18] used several ML models in order to predict the risk of liver disease. M. Suarez et al. [19] proposed a machine learning based technique for predicting liver disease. V. Harabor et al. [20] proposed a machine learning based approach for detecting hepatitis B and C.S. Tokala et al. [21] proposed several machine learning-based approaches for detecting liver problems. I.M. Attiya et al. [22] utilized supervised machine learning approaches for detecting liver problems. S.S. Nigatu et al. [23], D. Chicco & G. Jurman [24] and V.K. Yarasuri et al. [25] performed a comparison for hepatitis disease classification among several machine learning approaches. Their major focus was to improve classification model accuracy. Most of these studies delineated distinct methodologies and techniques instrumental in the diagnosis and prediction of Hepatitis. Diverse machine learning algorithms, encompassing Naive Bayes, Random Forest, SVM, KNN, Neural Networks, as well as Bagging and Boosting methods, were utilized in these studies. The contributions of each study were substantial, especially in terms of diagnostic accuracy and prediction precision.

3. Framework for Research and Related Resources

The framework has proposed in Figure 1 aims to develop an efficient hepatitis prediction model. In this research framework, a systematic approach is outlined for the development of an early Hepatitis detection model. It begins with data preprocessing, encompassing the handling of missing data using mean and mode imputation techniques as well as checking outliers using boxplot and z score. To solve the class imbalance problem, oversampling is applied. Algorithm selection involves the utilization of K-Nearest Neighbors (KNN), Naive Bayes (NB) and

Random Forest (RF), with the evaluation performed on balanced datasets. Feature selection is carried out using embedded feature selection technique to identify significant features. Performance evaluation includes assessing accuracy, precision, recall, f1 Score and roc. The ultimate aim is to construct a model for the early detection of hepatitis. To facilitate this research, relevant materials such as a trusted data source, programming in Python, and specific libraries as well as classification algorithms, missing data handling techniques, outliers removing techniques, class imbalance problem solving technique and feature selection methods play a vital role in executing the framework effectively.

3.1. Data collection

The dataset we utilized for this study is the Hepatitis Dataset sourced from the University of California,

Irvine's Machine Learning Repository. It is publicly available and can be accessed through [26]. This dataset provides a comprehensive view of factors related to hepatitis. It consists of 21 columns in total, encompassing various attributes related to patients who were diagnosed with hepatitis.

Below are the features and the target variable:

AGE: Age of the patient.
SEX: Gender of the patient (male or female).
STEROID: Indicates if the patient is on steroid treatment (yes or no).
ANTIVIRALS: Indicates if the patient is on antiviral treatment (yes or no).
FATIGUE: Presence of fatigue (yes or no).
MALAISE: Presence of malaise (yes or no).
ANOREXIA: Presence of anorexia (yes or no).
LIVER BIG: Enlarged liver (yes or no).
LIVER FIRM: Liver firm (yes or no).
SPLEEN PALPABLE: Palpable spleen (yes or no).
SPIDERS: Presence of spiders (yes or no).
ASCITES: Presence of ascites (yes or no).
VARICES: Presence of varices (yes or no).
BILIRUBIN: Bilirubin content in the blood.
ALK PHOSPHATE: Alkaline phosphatase content in the blood.
SGOT: SGOT content in the blood.
ALBUMIN: Albumin content in the blood.
PROTIME: Prothrombin time (in seconds).
HISTOLOGY: Presence of histological activity (yes or no).
TARGET (Class attribute): Indicates if the patient has hepatitis or not (live or die).

These attributes cover crucial numerical features such as age, levels of bilirubin, alkaline phosphate, SGOT, albumin, and protime, which are vital indicators for understanding the severity and progression of hepatitis. Additionally, the dataset includes nominal features like sex, steroid usage, presence of antivirals and information

about various symptoms such as fatigue, malaise, anorexia, liver and spleen conditions as well as the presence of ascites, varices and histology. One significant aspect to note is that this dataset contains missing values, implying that it requires careful preprocessing to handle these gaps appropriately and ensure the quality of the analysis. In terms of class distribution, the target variable "TARGET" signifies the presence or absence of hepatitis. The dataset exhibits an imbalance with 123 instances indicating the presence of hepatitis ("LIVE") where male is 16 and female 107 and 32 instances indicating the absence of hepatitis ("DIE") where male is 0 and female is 32. This class imbalance is a critical consideration for any analysis or modeling efforts, as it can influence the performance and interpretation of machine learning models.

3.2. Fundamental information and descriptive analysis

We have mentioned the working procedure of our study in Figure 1.

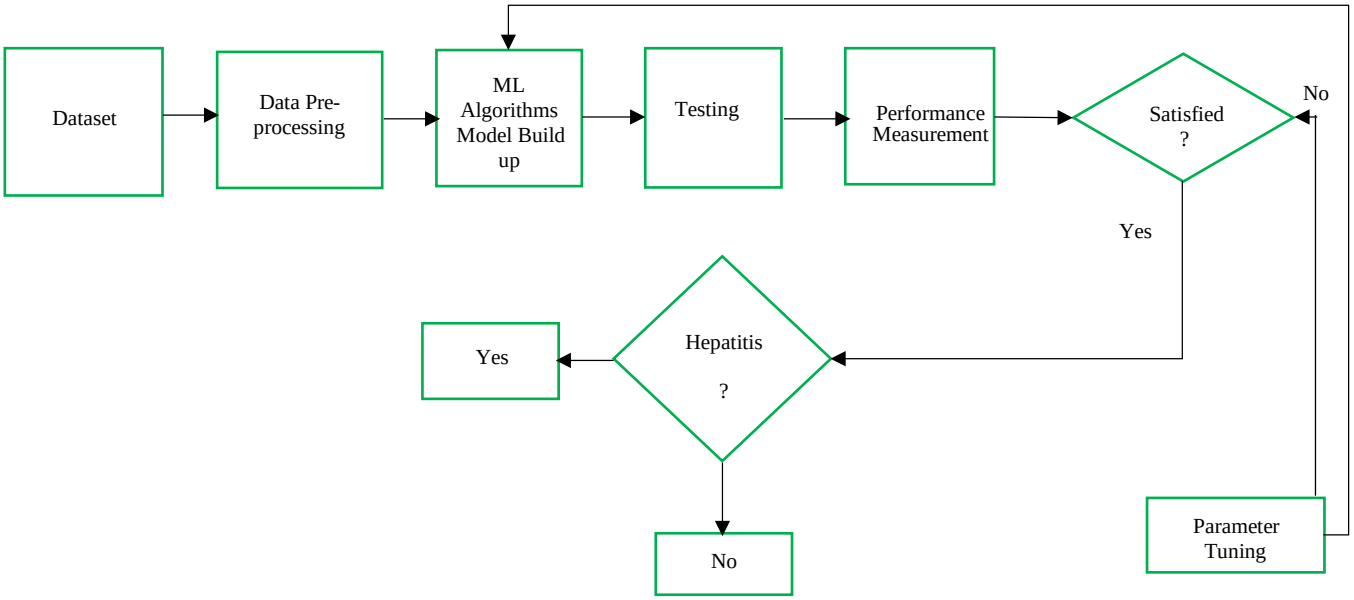


Figure 1. Working procedure of our study

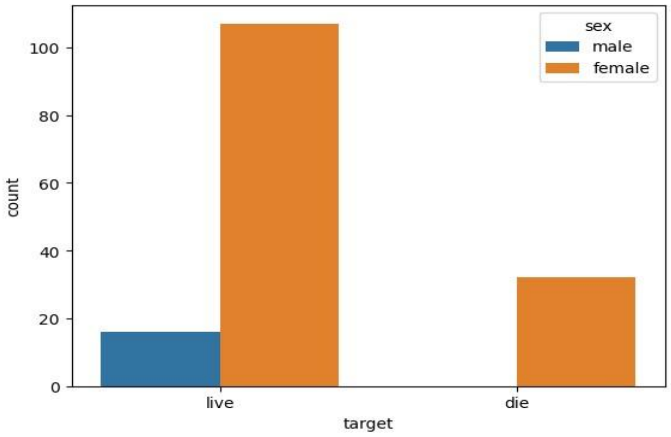


Figure 2. Gender Based Hepatitis Case Frequency Comparison for Original Dataset

The target class (live and die) is depicted visually in Figure 2 for both male and female in the Hepatitis original dataset. The graph provides a clear overview of the distribution of live and deceased cases among male and female patients, shedding light on the gender-specific outcomes within the context of Hepatitis. This visual depiction

allows for a quick understanding of the data, making it easier to discern any potential patterns or differences in disease outcomes between genders.

3.3. Data preprocessing

The comprehensive understanding of the dataset's statistical characteristics, coupled with effective handling of missing values and class imbalance, set the stage for a thorough evaluation and comparison of various classification models. Ultimately, this comprehensive research made it easier to choose the best models in order to predict hepatitis, optimizing their performance and ensuring reliable and accurate predictions. In the descriptive analysis of the dataset, key statistical measures were computed to provide a comprehensive understanding of the data's central tendencies, spread and distribution. These measures included the minimum, maximum, mean (average) and standard deviation (a measure of data dispersion) for each relevant attribute. These statistics were crucial in characterizing the dataset's range and variability, aiding in the interpretation of the dataset's features. From Table 1, we can notice some statistical result such as minimum (Min.), maximum (Max.), mean (Mean) and standard deviation (Std.) of several features of hepatitis disease. In the initial phase of data preprocessing, an exhaustive examination was conducted to identify inconsistencies and duplicates within the dataset. This involved scrutinizing the dataset for erroneous entries or repeated records to ensure data integrity and accuracy. Subsequently, a thorough analysis for missing values was performed, revealing specific columns with incomplete records. The mean and mode imputation techniques were applied to fill these missing values, calculated by averaging the available data points (mean) or identifying the most frequently occurring value (mode) in the respective columns.

Table 1. The hepatitis feature selection's statistical distribution (in pixels)

	Age	Bilirubin	Alk_phosphate	Sgot	Albumin	Protime
count	155.000000	155.000000	155.000000	155.000000	155.000000	155.000000
mean	41.200000	1.427517	105.325397	85.894040	3.817266	61.852273
std	12.565878	1.188301	46.405585	88.478932	0.616750	17.193528
min	7.000000	0.300000	26.000000	14.000000	2.100000	0.000000
25%	32.000000	0.800000	78.000000	32.500000	3.500000	57.000000
50%	39.000000	1.000000	102.000000	59.000000	3.900000	61.852273
75%	50.000000	1.500000	119.500000	99.000000	4.200000	65.000000
max	78.000000	8.000000	295.000000	648.000000	6.400000	100.000000

Outlier detection in Figure 3 and removal in Figure 4 was executed using the Z-score algorithm. The Z-score for each data point was calculated by measuring how many standard deviations it deviated from the mean. Data points exceeding a predefined threshold were identified as outliers and subsequently eliminated. Upon closer inspection, an imbalanced distribution of classes in the dataset was evident. To mitigate this issue, Oversampling technique was employed to balance the class representation. To further refine the dataset, Oversampling involved replicating instances of the minority class. This technique was accompanied by specific mathematical calculations to ensure a

balanced dataset. In recognizing the significance of relevant features, an embedded feature selection method was implemented. This technique involved assessing the importance of each feature in relation to the target variable, facilitating the selection of the most impactful features for model training and analysis. From figure 5, we can observe that the dataset after removed outliers exhibits an imbalance with 117 instances indicating the presence of hepatitis ("LIVE") where male is 15 and female 102 and 28 instances indicating the absence of hepatitis ("DIE") where male is 0 and female is 28. From figure 6, we can observe that the dataset after solved imbalance problem by oversampling technique exhibits a balance with 117 instances indicating the presence of hepatitis ("LIVE") where male is 15 and female 102 and 117 instances indicating the absence of hepatitis ("DIE") where male is 0 and female is 117. From figure 7, we can see the highly correlated features whose are present in our preprocessed dataset (free from outliers and imbalanced problem).

3.4. Classification model

In this study employed three key classification models: K-Nearest Neighbors (KNN), Naive Bayes (NB) and Random Forest (RF). These models were selected due to their relevance and effectiveness in predictive modeling for hepatitis. KNN is a proximity-based model, NB is a probabilistic model and RF is an ensemble learning model. Each of these models was meticulously applied and fine-tuned to extract meaningful insights and predictions related to hepatitis, contributing to a comprehensive understanding of the disease and aiding in early detection and intervention.

3.4.1. KNN

K-Nearest Neighbors (KNN) is a popular supervised machine learning algorithm and a non-parametric algorithm used for both classification and regression tasks to predict based on the similarity of the input instance to its k nearest neighbors in the training data.

Step-1: To process our algorithm, the value of K must be selected first. The process of choosing the right value of K is called parameter tuning. It is important for better accuracy. K value can be selected in two ways, one is heuristic in which we root on the total number of observations and avoid even values. Another is through experimentation in which we cross-validate the data set after dividing it into training and testing, then we need to build different models, the value of N will be different and all the different models should be trained on the training data, then the model that is trained should be run on the testing data. Then different accuracies will be available. Maximum accuracy will be selected.

Step-2: With the new value we have to find out the distance of all other data points from that point. It is done using the formula of Euclidian distance.

$$\text{Distance} = \sqrt{[(x2 - x1)2 + (y2 - y1)2]}$$

In step three, the data obtain from step two we will do the sorting data in ascending order. Step by step, we will take k number of values from the sorted data point then the class with the highest number is the class of that unknown value. According to the value of regression, we have to find out the mean of that number of values, the class that will come out, that is the class of the new value, here we will try to take 5 percent of the value of the total data set.

3.4.2. Naive bayes

Naive Bayes is a probabilistic machine learning algorithm based on Bayes' Theorem. It makes an assumption of feature independence given the class, which simplifies the probability calculations. The formula for Naive Bayes classification is as follows:

P (Ck|x1, x2,..., xn) = P(Ck)×P(x1|Ck) ×P(x2|Ck) ×…×P(xn|Ck)/ P(x1) × P(x2) ×…×P(xn) where:

P (Ck|x1, x2,..., xn) is the posterior probability of class, Ck given the features x1, x2,..., xn.

P(Ck) is the prior probability of class Ck.

P(xi|Ck) is the likelihood, the probability of feature xi given class Ck.

P(x1) ×P(x2) ×…×P(xn) is a scaling factor.

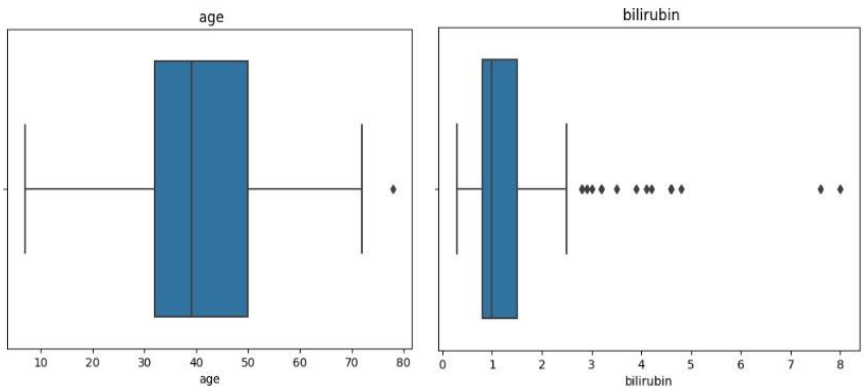
Naive Bayes assigns the class that maximizes this posterior probability for a given set of features.

3.4.3. Random forest

Bootstrapping the dataset and selecting random features are the two processes that make up Random Forest, a well-known ensemble learning technique. Bootstrapping is the process of creating numerous decision trees during training by randomly choosing subsets of the dataset with replacement. Each of these subsets is employed to train a decision tree. Moreover, at each node of these trees, only a random subset of features is considered for splitting. This strategy enhances diversity and reduces correlation among the trees within the forest. When it comes to making predictions, each tree in the forest predicts a class for a new input. When it comes to classification purposes, the predictable classes from each individual tree are voted on in a majority manner to decide the final result. Specifically, for a new sample, the predicted class C is computed as the mode of the predicted classes ($C1$, $C2$,..., CB) from each tree in the forest, where B is the number of trees. In mathematical terms, for classification, the predicted class C for a new sample is given by:

C=mode ($C1$, $C2$,..., CB)

where Ci represents the predicted class by the i-th tree in the forest. Essentially, each tree contributes a prediction and the final prediction is determined by the most commonly predicted class, achieving a robust and reliable classification outcome. On the other hand, for regression tasks, the final prediction is often computed as the mean of the predictions from the individual trees.



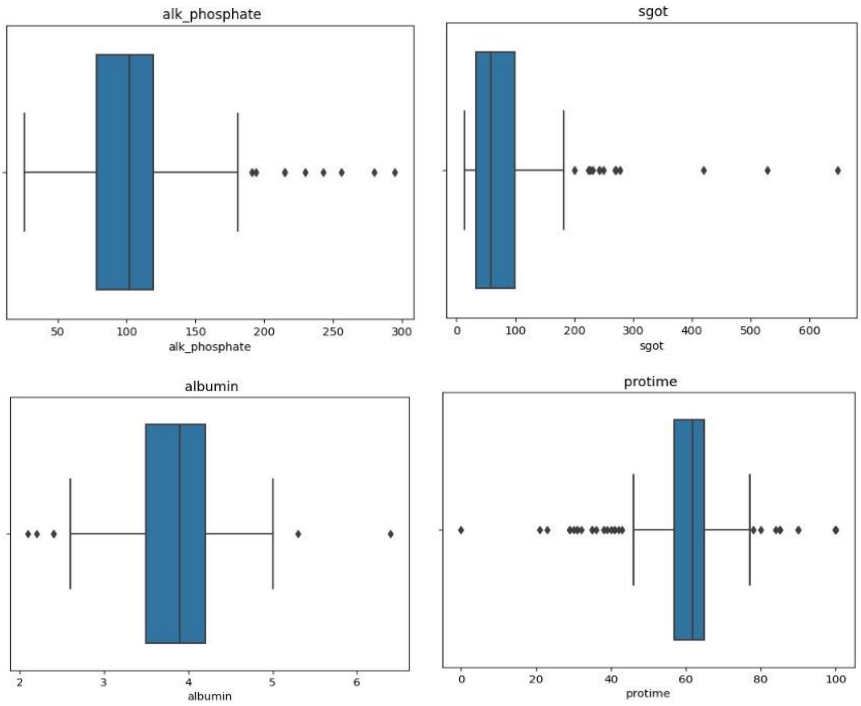


Figure 3. Boxplot for Various Hepatitis Features

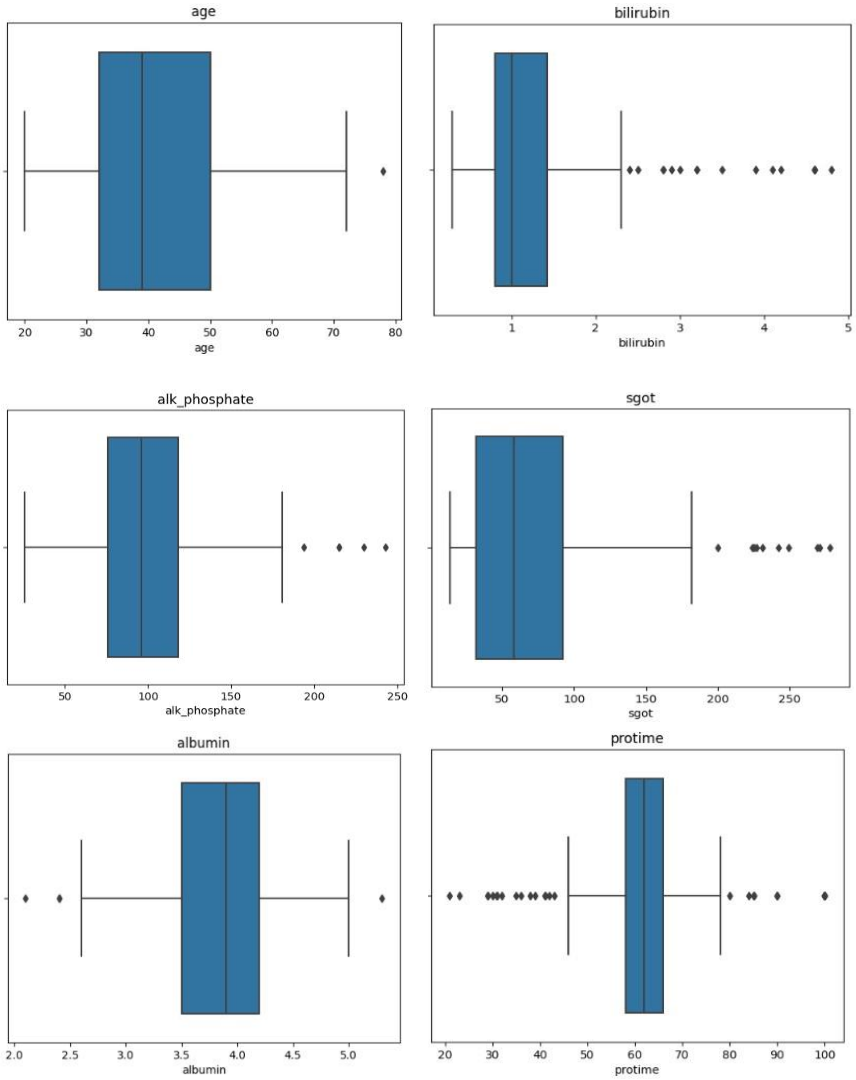


Figure 4. Boxplot for Various Hepatitis Features (After Removing Outliers)

3.5. Performance measures

In assessing our hepatitis prediction models, a range of performance metrics was employed. Accuracy gauged overall correctness, while precision emphasized true positive rates. Recall focused on true positive recovery, vital for illness identification. The F1 score struck a balance between precision and recall, capturing model effectiveness. The ROC-AUC score evaluated the model's discrimination ability effectively. Specificity highlighted true negative identification, crucial in hepatitis prediction. Sensitivity underscored true positive detection, a pivotal aspect of model assessment. These metrics collectively provided a comprehensive evaluation of our models' predictive capabilities for hepatitis. Table 2 displays the binary class confusion matrix, providing a visual representation of the predicted and actual class values. In the context of classification, this matrix is an invaluable tool for evaluating each class's performance. The evaluation of accurate predictions for classification problems is achieved through a comprehensive set of performance metrics, as detailed in Table 3. These metrics encompass Accuracy (ACC), Precision, Recall, F1 Score True Positive Rate, False Positive Rate (FPR). In our research, we have applied 10-fold cross validation in order to calculate different parameters by which evaluated our proposed models' performances. In our research, we have utilized python programming language (version: 3.12.1) in order to build up our proposed models and evaluate their performances.

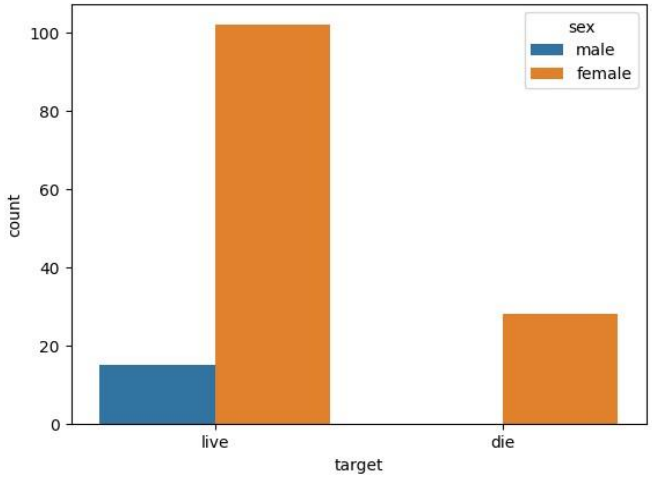


Figure 5. Gender Based Hepatitis Case Frequency Comparison after removing outliers

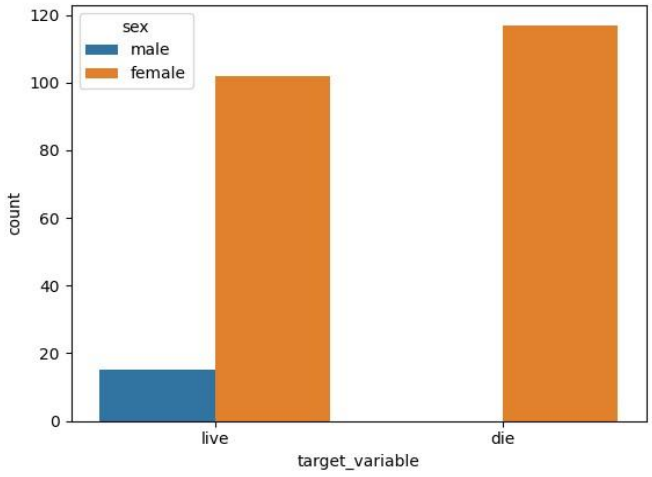


Figure 6. Gender Based Hepatitis Case Frequency Comparison after solved imbalance problem

In Figure 7, the embedded feature selection method effectively identifies and highlights the most crucial features from a myriad of available attributes. It assigns scores to each feature, indicating their significance in accurately classifying hepatitis. These important feature scores serve as a valuable guide for selecting the most relevant attributes, optimizing the predictive model, and enhancing the overall performance of hepatitis classification. We have got total sixteen highly correlated features among eighteen independent features by applying embedded feature selection technique. The highly correlated independent features are: age, sex, steroid, fatigue, malaise, anorexia, spleen_palpable, spiders, ascites, varices, bilirubin, alk_phosphate, sgot, albumin, protime and histology. In our final preprocessed dataset, there are total seventeen features (sixteen highly correlated independent features and one dependent target or predictable feature).

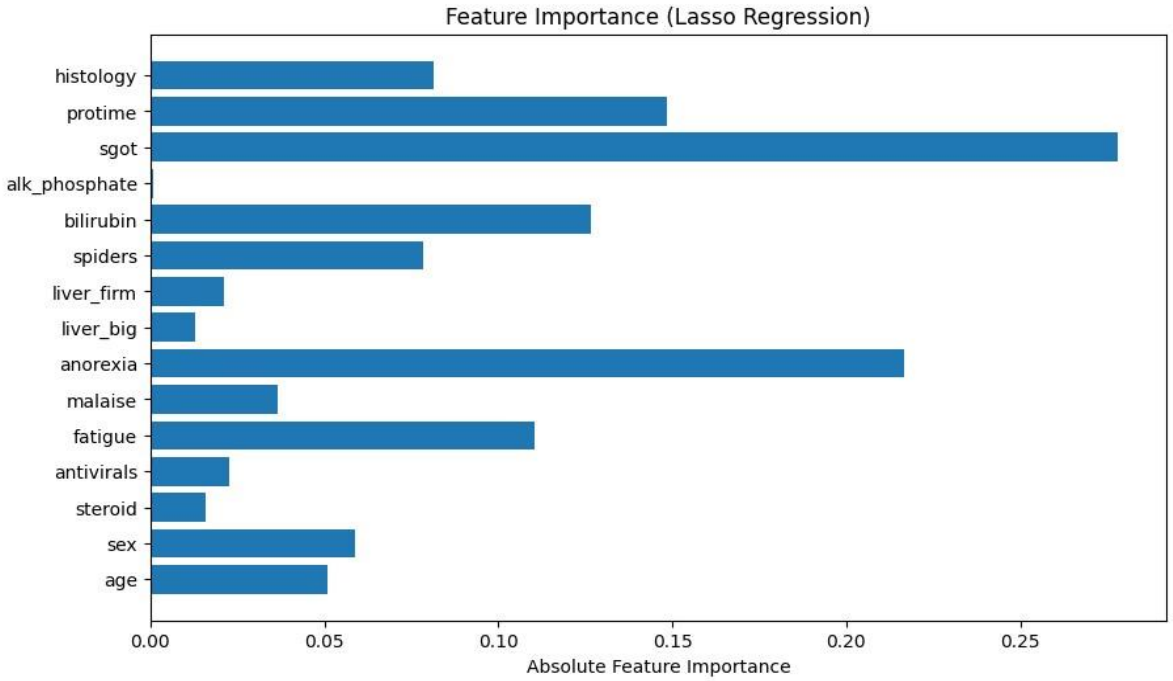


Figure 7. Feature importance by using embedded feature selection technique

Table 2. Confusion matrix and representation for multiple classes

	Predicted Positive	**Predicted Negative**
Actual Positive	True Negative (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Positive (TN)

Table 3. Essential formulas together with descriptions for assessing our models

Name of the Formula	**Equation**	**Description**
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	Total correctness of prediction.
Precision	$\frac{TP}{TP+FP}$	Relevance of positive prediction.

Recall	$\frac{TP}{TP+FN}$	Models' ability to capture all positives.
F1-Score	$2* \frac{precision*recall}{precision+recall}$	Balance between precision and recall.
True Positive Rate (TPR)	$\frac{TP}{TP+FN}$	The probability of an actual positive data will test positive.
False Positive Rate (FPR)	$\frac{TN}{TN+FP}$	The probability of an actual negative data will test negative

3.5.1. Receiver operating characteristics (ROC) curve

An essential tool for assessing our hepatitis prediction models was the ROC curve. This graphical depiction shows how the model can distinguish between different classes. At different thresholds criteria, the True Positive Rate is shown against the False Positive Rate in a curve. The area under this curve (AUC-ROC) quantifies the model's discriminatory power. A higher AUC-ROC suggests better performance, aiding in selecting the optimal model for hepatitis prediction. The ROC curve and AUCROC were pivotal in our assessment, ensuring the accuracy and effectiveness of our predictive models.

4. Experimental Results

We have initially focused on the confusion matrix and our experimental settings in this section. After that, move on to a thorough examination and discussion of the outcomes that were produced as well as the efficacy of the suggested framework, examining these aspects sequentially.

4.1. Confusion matrix

Understanding how well a model performs in terms of accuracy, precision, recall, F1 Score and other performance metrics requires looking at the confusion matrix, which offers a clear split of these four outcomes. It is a pivotal tool in assessing the quality of a classification algorithm and is often used in conjunction with these metrics to gain insights into the model's effectiveness.

4.2. Experimental different types of results

Table 4. Performance of our proposed models after solving missing value problems

Algorithm	Confusion Matrix		Accuracy	Precision	Recall	F1 Score
KNN	TP = 109	FN = 14	81.29%	0.88	0.88	0.88
	FP = 15	TN = 17				
Naive Bayes	TP = 102	FN = 21	75.48%	0.86	0.83	0.84
	FP = 17	TN = 15				
Random Forest	TP = 118	FN = 5	**87%**	**0.89**	**0.96**	**0.92**
	FP = 15	TN = 17				

Table 5. Performance of our proposed models after removing missing value and outliers and imbalanced problem

Algorithm	Confusion Matrix		Accuracy	Precision	Recall	F1 Score
KNN	TP = 92	FN = 24	86.32%	0.92	0.79	0.85
	FP = 8	TN = 110				
Naive Bayes	TP = 97	FN = 20	80.34%	0.79	0.82	0.80
	FP = 26	TN = 91				
Random Forest	TP = 106	FN = 10	**95.30%**	**0.99**	**0.91**	**0.95**
	FP = 1	TN = 117				

Table 6. Performance of our proposed models after removing missing value and outliers and solved imbalanced problem and incorporated highly correlated features

Algorithm	Confusion Matrix		Accuracy	Precision	Recall	F1 Score
KNN	TP = 88	FN = 28	83.33%	0.89	0.75	0.81
	FP = 11	TN = 107				
Naive Bayes	TP = 89	FN = 28	73.50%	0.72	0.76	0.75
	FP = 34	TN = 83				
Random Forest	TP = 109	FN = 5	**97.44%**	**0.99**	**0.96**	**0.97**
	FP = 1	TN = 119				

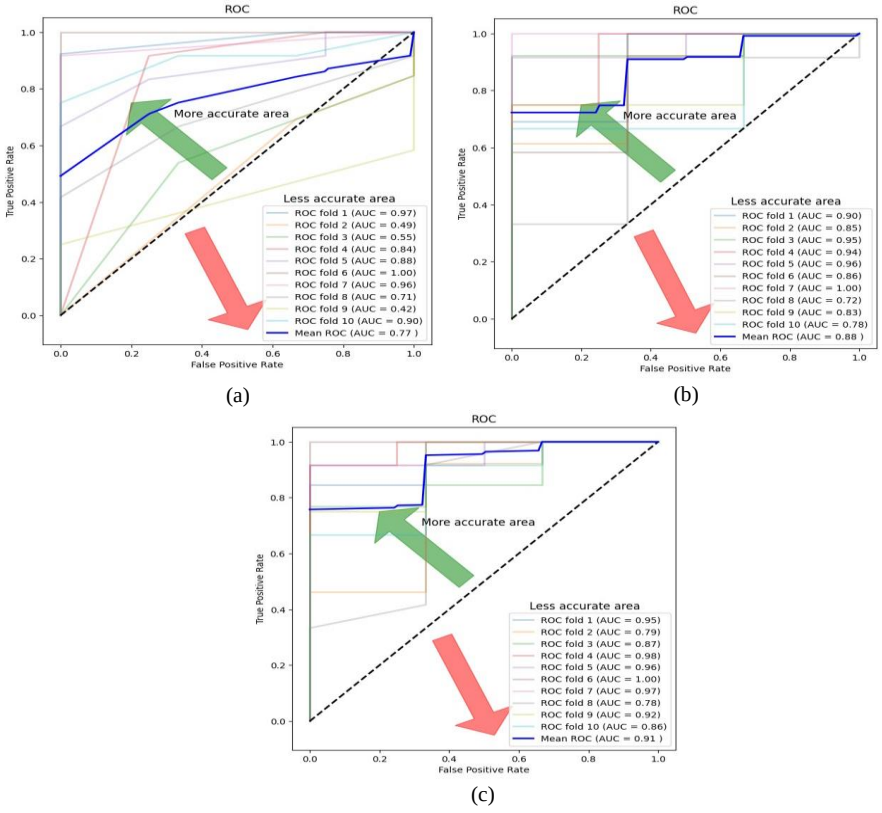


Figure 8. ROC Curve Analysis of (a) KNN, (b) NB and (c) RF models after Mean/Mode Imputation

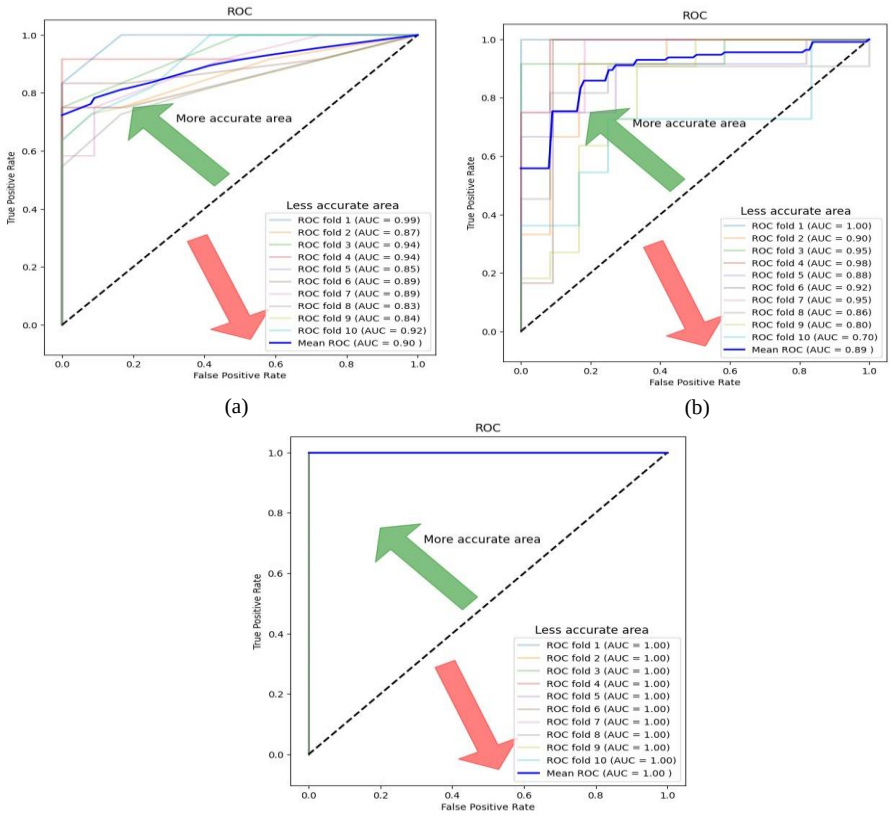


Figure 9. ROC Curve Comparison of (a) KNN, (b) NB and (c) RF models after Mean/Mode Imputation, removed outliers and incorporated Oversampling technique to solve imbalance problem.

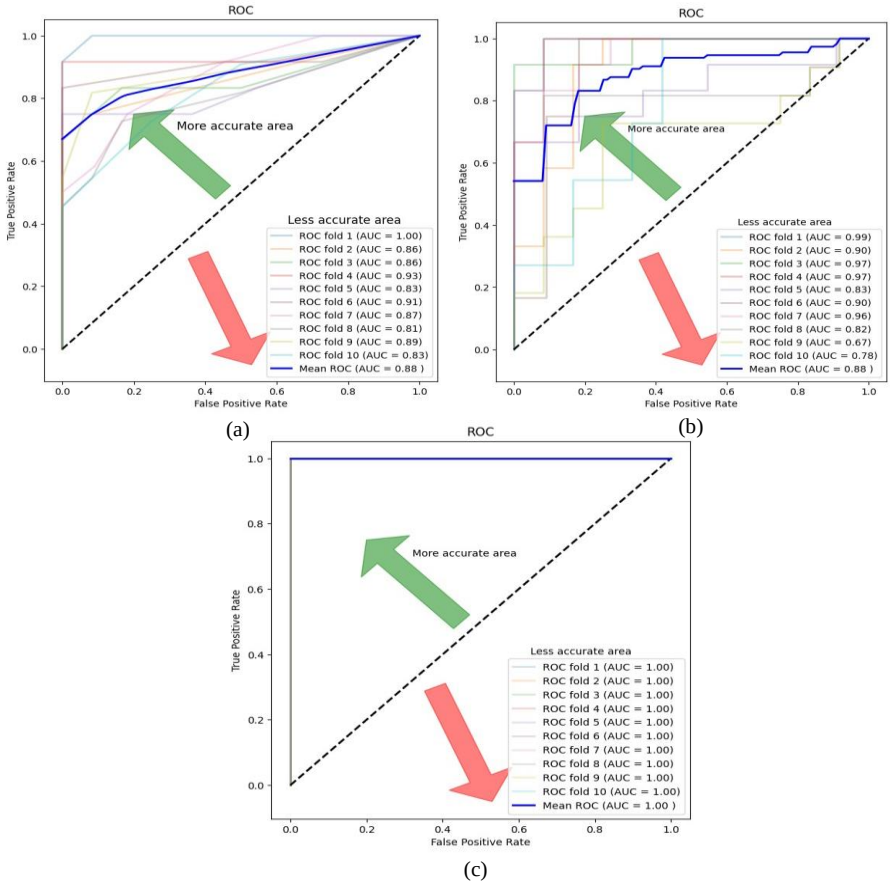


Figure 10. ROC Curve Comparison of (a) KNN, (b) NB and (c) RF models after Mean/Mode

Imputation, removed outliers, solved imbalance problem by Oversampling and incorporated Embedded Feature Selection Technique.

5. Discussion

In this framework, the primary focus was on leveraging advanced machine learning techniques to effectively classify hepatitis cases, an extensive analysis of a dataset was conducted, laying the foundation for accurate predictive modeling and insightful analysis.

Table 7. Comparing the performance with existing works

Research	**Published Year**	**Testing Data**	**Selected Models**	**Best Model**	**Accuracy (%)**
Krisnabayu et al. [10]	2021	Hepatitis	RF	RF	87.9%
Butt et al. [6]	2021	Hepatitis C	Artificial Back Propagation	Artificial Back Propagation	94.44%
Nayeem et al. [5]	2021	Hepatitis	KNN, NB, SVM, RF, Multilayer Perceptron	RF	92.41%
Gundogdu [3]	2022	Hepatitis C	PCA-SVM	PCA-SVM	95.7%
Farghaly et al. [2]	2023	Hepatitis C	KNN, NB, LR, RF	RF	94.88%
Dutta et al. [11]	2023	Hepatitis	RF	RF	93.91%
Sachdeva et al. [12]	2023	Hepatitis	SVM, LR, KNN, RF	LR	93.81%
Proposed Model	-	Hepatitis	KNN, NB, RF	**RF**	**97.44%**

The preprocessing phase was meticulous, starting with a rigorous check for inconsistencies and duplicates, followed by the handling of missing values using mean and mode techniques. Addressing class imbalances through oversampling was crucial for unbiased model training. The classification models selected, including K-Nearest Neighbors (KNN), Naive Bayes (NB) and Random Forest (RF), were chosen for their suitability and effectiveness in predictive modeling for hepatitis. Taking the data preprocessing a step further, an embedded feature selection method was introduced. This step refined the dataset, leading to a further boost in model performance and accuracy. The models were thoroughly trained and evaluated on the preprocessed, oversampled and highly correlated feature selected datasets. From Fig-8, fig-9 and fig-10, It can be noticed that Random Forest exhibited superior performance, especially after oversampling was applied, showcasing the significance of dealing with class imbalances. From table 4, table 5 and table 6, It can be noticed that the highest classification accuracy has achieved from Random Forest algorithm for all cases. Accurate predictive models can greatly assist in early detection and timely intervention for individuals at risk of hepatitis. This approach equalized the distribution of features across all classes, enhancing classification accuracy. Performance evaluation involved key metrics, including Accuracy (ACC), precision, recall, F1 Score and ROC (AUC) providing insights into model robustness. Among the considered classification models, RF exhibited superior performance on the balanced with the existence of highly correlated features in our hepatitis dataset, with high accuracy, precision, recall and ROC (AUC). Figure 8, 9 and 10 clearly demonstrate the notable differences in ROC values obtained through various dataset preprocessing

techniques. Their compelling visual representation highlights the efficacy of these methods in achieving significantly higher ROC values. Upon comparing the results presented in Figure 8, 9 and 10, it becomes evident that the Random Forest (RF) model, coupled with embedded feature selection, exhibits the highest level of accuracy and it is 96.44%.

5.1. Causes for the highest performance comes from random forest

The ML algorithm with the RF classifier demonstrated superior performance attributed to its ensemble learning nature and specific strategies. RF aggregates predictions from multiple decision trees, enhancing accuracy and resilience against overfitting. Random feature selection and bootstrapped sampling add diversity and prevent overfitting. Additionally, strategic feature selection, notably using information gain and embedded techniques, pinpointed vital features, amplifying the classifier's accuracy and robustness in predicting hepatitis. The combined power of ensemble learning and thoughtful feature selection made RF a key player in this hepatitis classification study.

5.2. Strengths and drawbacks of the research

In our study, achieving early hepatitis prediction with high accuracy stands as a significant strength. By utilizing advanced machine learning models and carefully preprocessing the dataset, we were able to provide timely predictions for hepatitis presence. This is critical for prompt medical intervention and timely treatment, potentially preventing the progression of the disease and improving patient outcomes. However, there are limitations to our study. The dataset, though extensive, may not encompass all possible scenarios related to hepatitis. Additionally, while we handled missing values and imbalanced data, there might still be inherent biases impacting the model predictions. Further research with larger and more diverse datasets is necessary to mitigate these limitations.

6. Conclusion

In this study, we delved into the crucial domain of hepatitis classification with a keen eye on imbalanced and balanced dataset scenarios. This holds immense significance in both medical research and practical healthcare applications. The accuracy and early prediction of hepatitis are vital in ensuring timely intervention and effective disease management. We employed a diverse dataset related to hepatitis, employing various machine-learning algorithms and techniques. Among them, Random Forest (RF) showed exceptional promise in handling imbalanced datasets, providing accurate predictions with Mean Mode Imputation 87%, with RF with Mean, Mode Imputation, outlier removed & oversampling 95.30%, with Mean, Mode Imputation, outlier removed & oversampling and also with embedded system 97.44%. The feature selection techniques revealed the pivotal attributes contributing to hepatitis prediction. Specifically, our findings highlighted that early prediction of hepatitis can be achieved with substantial accuracy using the Random Forest algorithm. This holds significant promise for the medical community and healthcare practitioners in devising proactive strategies for hepatitis management. Additionally, we acknowledged certain limitations in our study, such as the need for a more extensive dataset and the necessity of exploring other advanced machine learning algorithms. These factors suggest promising avenues for future research in hepatitis classification, potentially leading to more accurate and efficient predictive models. Future research should focus on improving the accuracy of early hepatitis prediction while

maintaining timely results. Incorporating more diverse and comprehensive data, refining feature selection techniques, advanced machine learning algorithms and exploring deep learning techniques. Additionally, integrating more extensive medical data and advanced imaging technologies could lead to more precise predictions. Besides this, collaboration with medical experts to validate and fine-tune the models and collect large datasets will be crucial in achieving highly accurate and early predictions for hepatitis.

Declarations

Source of Funding

This study did not receive any grant from funding agencies in the public, commercial, or not–for–profit sectors.

Competing Interests Statement

The authors declare no competing financial, professional, or personal interests.

Consent for publication

The authors declare that they consented to the publication of this study.

Authors' contributions

All the authors made an equal contribution in the Conception and design of the work, Data collection, Simulation analysis, Drafting the article, and Critical revision of the article. All the authors have read and approved the final copy of the manuscript.

Availability of data and material

Authors are willing to share data and material according to the relevant needs.

References

[1] World Health Organization: WHO (2020). Hepatitis. https://www.who.int/health-topics/hepatitis#tab=tab_1.

[2] Farghaly, H.M, Shams, M.Y., & Abd El-Hafeez, T. (2023). Hepatitis C Virus prediction based on machine learning framework: a real-world case study in Egypt. Knowledge and Information Systems, 65(6): 2595–2617. https://doi.org/10.1007/s10115-023-01851-4.

[3] Gündoğdu, S. (2022). Hepatitis C Disease Detection Based on PCA–SVM Model. Hittite Journal of Science and Engineering, 9(2): 111–116. https://doi.org/10.17350/hjse19030000261.

[4] Majzoobi, M.M., Namdar, S., Najafi-Vosough, R., Hajilooi, A.A., & Mahjub, H. (2022). Prediction of Hepatitis disease using ensemble learning methods. Journal of Preventive Medicine and Hygiene, 63(3). https://doi.org/10.15167/2421-4248/jpmh2022.63.3.2515.

[5] Nayeem, M.J., Rana, S., Alam, F., & Rahman, M.A. (2021). Prediction of hepatitis disease using K-nearest neighbors, Naive Bayes, support vector machine, multi-layer perceptron and random forest. IEEE International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD), Pages 280–284. https://doi.org/10.1109/icict4sd50815.2021.9397013.

[6] Butt, M.B., Alfayad, M., Saqib, S., Khan, M.A., Ahmad, M., Khan, M.A., & Elmitwally, N.S. (2021). Diagnosing the stage of hepatitis C using machine learning. Journal of Healthcare Engineering, Pages 1–8. https://doi.org/10.1155/2021/8062410.

[7] Hafeez, M.A., Imran, A., Khan, M.I., Khan, A.H., Nawaz, A., & Ahmed, S. (2022). Diagnosis of Liver Disease Induced by Hepatitis Virus Using Machine Learning Methods. IEEE 8th International Conference on Information Technology Trends (ITT), Pages 154–159. http://dx.doi.org/10.1109/itt56123.2022.9863944.

[8] Trishna, T.I., Emon, S.U., Ema, R.R., Sajal, G.I.H., Kundu, S., & Islam, T. (2019). Detection of hepatitis (a, b, c and e) viruses based on random forest, k-nearest and naïve bayes classifier. IEEE 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Pages 1–7. https://doi.org/10.1109/icccnt45670.2019.8944455.

[9] Hashem, S., Esmat, G., Elakel, W., Habashy, S., Abdel Raouf, S., Darweesh, S., & ElHefnawi, M. (2016). Accurate prediction of advanced liver fibrosis using the decision tree learning algorithm in chronic hepatitis C Egyptian patients. Gastroenterology Research and Practice, Pages 1–7. https://doi.org/10.1155/2016/2636390.

[10] Krisnabayu, R.Y., Ridok, A., & Setia Budi, A. (2021). Hepatitis detection using random forest based on SVM-RFE (recursive feature elimination) feature selection and SMOTE. In Proceedings of the 6th International Conference on Sustainable Information Engineering and Technology, Pages 151–156. https://doi.org/10.1145/3479645.3479668.

[11] Dutta, P., Paul, S., Jana, G.G., & Sadhu, A. (2023). Hybrid Genetic Algorithm Random Forest algorithm (HGARF) for improving the missing value Imputation in Hepatitis Medical Dataset. IEEE International Symposium on Devices, Circuits and Systems (ISDCS), Pages 01–05. https://doi.org/10.1109/isdcs58735.2023.10153526.

[12] Sachdeva, R.K., Bathla, P., Rani, P., Solanki, V., & Ahuja, R. (2023). A systematic method for diagnosis of hepatitis disease using machine learning. Innovations in Systems and Software Engineering, 19(1): 71–80. https://doi.org/10.1007/s11334-022-00509-8.

[13] Genemo, M.D. (2023). Diagnosis of Hepatitis using Supervised Learning algorithm. Indonesian Journal of Data and Science (IJODAS), 4(1): 25–30. https://doi.org/10.56705/ijodas.v4i1.60.

[14] Ahmed, I.I., Mohammed, D.Y., & Zidan, K.A. (2022). Diagnosis of hepatitis disease using machine learning techniques. Indonesian Journal of Electrical Engineering and Computer Science, 26(3): 1564–1572. http://doi.org/10.11591/ijeecs.v26.i3.pp1564-1572.

[15] Alizargar, A., Chang, Y.L., & Tan, T.H. (2023). Performance comparison of machine learning approaches on Hepatitis C prediction employing data mining techniques. Bioengineering, 10(4): 481. https://doi.org/10.3390/bioengineering10040481.

[16] Md, A.Q., Kulkarni, S., Joshua, C.J., Vaichole, T., Mohan, S., & Iwendi, C. (2023). Enhanced Preprocessing Approach Using Ensemble Machine Learning Algorithms for Detecting Liver Disease. Biomedicines, 11(2): 581. https://doi.org/10.3390/biomedicines11020581.

[17] Alotaibi, A., Alnajrani, L., Alsheikh, N., Alanazy, A., Alshammasi, S., Almusairii, M., & Alansari, A. (2023). Explainable Ensemble-Based Machine Learning Models for Detecting the Presence of Cirrhosis in Hepatitis C Patients. Computation, 11(6): 104. https://doi.org/10.3390/computation11060104.

[18] Dritsas, E., & Trigka, M. (2023). Supervised machine learning models for liver disease risk prediction. Computers, 12(1): 19. https://doi.org/10.3390/computers12010019.

[19] Suárez, M., Martínez, R., Torres, A.M., Ramón, A., Blasco, P., & Mateo, J. (2023). A Machine Learning-Based Method for Detecting Liver Fibrosis. Diagnostics, 13(18): 2952. https://doi.org/10.3390/diagnostics 13182952.

[20] Harabor, V., Mogos, R., Nechita, A., Adam, A.M., Adam, G., Melinte-Popescu, A.S., & Harabor, A. (2023). Machine Learning Approaches for the Prediction of Hepatitis B and C Seropositivity. International Journal of Environmental Research and Public Health, 20(3): 2380. https://doi.org/10.3390/ijerph20032380.

[21] Tokala, S., Hajarathaiah, K., Gunda, S.R.P., Botla, S., Nalluri, L., Nagamanohar, P., & Enduri, M.K. (2023). Liver Disease Prediction and Classification using Machine Learning Techniques. International Journal of Advanced Computer Science and Applications, 14(2): 1–9. http://dx.doi.org/10.14569/ijacsa.2023.0140299.

[22] Attiya, I.M., Abouelsoud, R.A., & Ismail, A.S. (2023). A Proposed Approach for Predicting Liver Disease. Information Sciences Letters, 12(6): 2447–2460. http://dx.doi.org/10.18576/isl/120644.

[23] Nigatu, S.S., Alla, P.C.R., Ravikumar, R.N., Mishra, K., Komala, G., & Chami, G.R. (2023). A Comparative Study on Liver Disease Prediction using Supervised Learning Algorithms with Hyperparameter Tuning. IEEE International Conference on Advancement in Computation & Computer Technologies (InCACCT), Pages 353–357. https://doi.org/10.1109/incacct57535.2023.10141830.

[24] Chicco, D., & Jurman, G. (2021). An ensemble learning approach for enhanced classification of patients with hepatitis and cirrhosis. IEEE Access, 9: 24485–24498. https://doi.org/10.1109/access.2021.3057196.

[25] Yarasuri, V.K., Indukuri, G.K., & Nair, A.K. (2019). Prediction of hepatitis disease using machine learning technique. IEEE 3rd International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), Pages 265–269. https://doi.org/10.1109/i-smac47947.2019.9032585.

[26] Hepatitis (1988). UCI Machine Learning Repository. https://doi.org/10.24432/c5q59j.